Caracterização do Comportamento dos Clientes de Vídeo ao Vivo Durante um Evento de Larga Escala

Breno Santos[1] **Gustavo Carnivali**[1] **Wagner Almeida**[1,3] **Alex B. Vieira**[1]
Ítalo Cunha[2] **Jussara Almeida**[2]

[1]Departamento de Ciência da Computação, Universidade Federal de Juiz de Fora
[2]Departamento de Ciência da Computação, Universidade Federal de Minas Gerais
[3]Instituto Federal do Sudeste de Minas Gerais

{wagner.almeida, breno.santos, gustavocarnivali}@ice.ufjf.br {alex.borges}@ufjf.edu.br
{cunha, jussara}@dcc.ufmg.br

Abstract. *Internet large-scale events are no longer an exception. Although, both researchers and service providers still have limited visibility and understanding about these events. In this paper, we present an in-depth live streaming client behavior characterization during a large scale event. We analyze 64 soccer matches data from FIFA's 2014 Soccer World Cup collected at servers from one of the major Internet content providers in Brazil. We have identified the main client behavior features and propose a simple model that captures its behavior. Our results are compared to the state-of-the-art and indicate a change in the client behavior during Internet large-scale events. We believe that our results can be applied to generate realistic synthetic workloads and serve as a substrate for the development and evaluation of new Internet live streaming architectures.*

Resumo. *Eventos de larga escala na Internet não mais são exceções. Entretanto, tanto pesquisadores quanto provedores de serviço ainda têm visão e entendimento limitados sobre tais eventos. Neste artigo, nós caracterizamos o comportamento dos usuários de um sistema de vídeo ao vivo durante um grande evento na Internet, a copa do mundo de futebol da FIFA em 2014. Analisamos dados relativos à transmissão de 64 partidas por um dos principais provedores de conteúdo Internet do Brasil. Nós identificamos as principais características do comportamento de um cliente e propomos um modelo simples que captura seu comportamento típico. Nossos resultados são comparados ao estado da arte e indicam mudanças no comportamento dos clientes durante eventos de larga escala na Internet. Acreditamos que nossos resultados podem ser aplicados na geração de cargas sintéticas realistas e servir como substrato para o desenvolvimento e avaliação de novas arquiteturas de transmissão ao vivo na Internet.*

1. Introdução

Eventos em larga escala são uma realidade cada vez mais comum na Internet. De fato, desde o discurso de posse de Obama[1], provedores de serviços estão sujeitos a surtos no número de clientes. Infelizmente, os provedores de serviços são incapazes de prever com precisão a demanda desses eventos, particularmente para eventos esporádicos ou únicos, frequentemente subestimando ou superestimando a demanda. Provedores de serviço transmitem conteúdo de infraestruturas baseadas em nuvem (como centros de processamento de dados ou redes de distribuição de conteúdo) [Barroso et al. 2013]. Neste

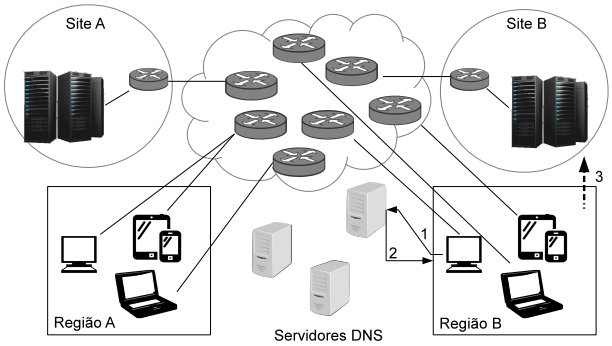


Figura 1. Arquitetura de distribuição de vídeo ao vivo do sistema estudado.

cenário, subestimar a demanda resulta em serviço de baixa qualidade e insatisfação dos clientes, enquanto superestimar a demanda implica custos adicionais de operação.

Apesar dos trabalhos de caracterização de transmissão de mídia ao vivo na existentes [Veloso et al. 2006, Hei et al. 2007, Borges et al. 2012], pesquisadores e provedores de serviço ainda têm entendimento limitado sobre eventos de larga escala na Internet. De fato, enquanto os serviços na Internet evoluem, o padrão de comportamento de seus clientes também muda [Gebert et al. 2012]. Consequentemente, indústria e academia necessitam de caracterizações e modelos atualizados, de modo que possam desenvolver aplicações mais complexas e arquiteturas de serviços mais eficientes.

Nesse cenário, nós apresentamos uma caracterização do comportamento dos usuários de um sistema de vídeo ao vivo durante a copa do mundo de futebol da FIFA em 2014 (seção 3). Nós analisamos carga de trabalho da transmissão das 64 partidas de futebol coletada em servidores de um dos principais provedores de conteúdo do Brasil (seção 2.2). Nós identificamos as principais características do comportamento de clientes e propomos um modelo simples que captura o comportamento típico de um cliente desse tipo de aplicação (seção 3.1). Nós geramos uma caracterização detalhada, permitindo assim geração de carga sintética realista. Nós contrastamos o comportamento observado durante a copa do mundo de 2014 com o comportamento observado em caracterizações de outros eventos transmitidos ao vivo pela Internet. Por último, nós desenvolvemos um gerador de carga de cargas sintéticas capaz de capturar, em momentos estáveis, o comportamento dos clientes no sistema estudado.

2. Sistema de distribuição de mídia ao vivo e conjunto de dados

2.1. Componentes do sistema

A infraestrutura de transmissão de vídeo ao vivo do provedor de serviço estudado possui dois pontos de distribuição com diversos servidores. Esses pontos estão localizados em duas das mais maiores cidades do Brasil, Rio de Janeiro e São Paulo. Como mostramos na Figura 1, em cada um desses pontos de distribuição, o provedor de conteúdo é conectado a um ponto de troca de tráfego (PTT) local e a várias redes comerciais.

O sistema de transmissão de vídeo ao vivo do provedor de serviço estudado usa *anycast* [Cesario 2012], uma técnica de engenharia de tráfego onde um prefixo IP é anun-

[1] https://gigaom.com/2009/02/07/cnn-inauguration-p2p-stream-a-success-despite-backlash/

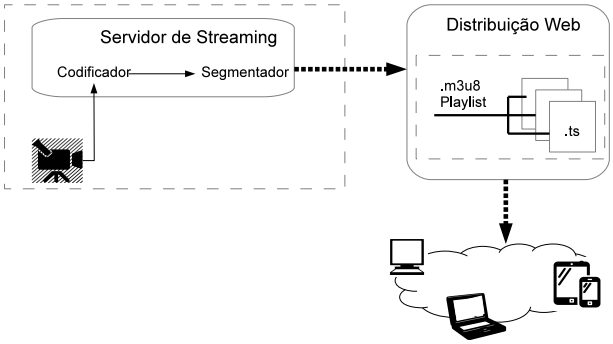


Figura 2. Esquema de codificação e distribuição do conteúdo ao vivo.

ciado a partir de múltiplos locais. A rede propaga os anúncios e protocolos de roteamento escolhem as melhores rotas. As rotas escolhidas, em geral, conectam clientes ao servidor geograficamente mais próximo, levando a menor latência, maior banda, e distribuição de carga [Katabi and Wroclawski 2000]. Assim, como ilustramos na Figura 1, um cliente acessa o domínio do provedor de conteúdo a partir de seu nome (1). O nome de domínio é resolvido por um servidor de DNS apropriado (2). O IP que o servidor de DNS resolveu corresponde a um IP anunciado via *anycast* que pode ser encaminhado a qualquer ponto de distribuição (3). Note que *anycast* divide os clientes entre os dois pontos de distribuição, mas não garante balanceamento.

Atualmente, a maioria dos servidores de envio de mídia, por exemplo, YouTube e Netflix, entregam conteúdo usando o HTTP. O HTTP é ubíquo e amplamente disponível, superando empecilhos comumente encontrados por sistemas de distribuição de mídia ao vivo que utilizando comunicação par-a-par (P2P). Em particular, transmissão por HTTP requer somente um navegador Web para visualizar conteúdos de vídeo e é desnecessário abrir portas em *firewalls* ou configurar redirecionamento de conexões externas em tradutores de endereços de rede (NATs).

O provedor de conteúdo estudado disponibiliza a mídia ao vivo em múltiplas taxas de codificação, com diferentes níveis de qualidade, usando *HTTP Live Streaming* (HLS). Um sistema com HLS funciona da seguinte maneira (Figura 2): inicialmente, o vídeo é capturado e codificado em várias taxas de transmissão (qualidade) diferentes. Cada mídia codificada é dividida em *segmentos*. Pequenos lotes de segmento com a mesma taxa de codificação são agrupados e indexados em *listas de reprodução*. A lista de reprodução permite ao cliente requisitar segmentos e reproduzir um trecho do vídeo.

Para receber o conteúdo ao vivo, cada cliente estima qual taxa de transmissão do vídeo ao vivo sua banda de rede suporta. Então, o cliente solicita a lista de reprodução referente a essa qualidade (i.e., um arquivo com extensão m3u8). Ao fim da reprodução dos segmentos de vídeo em uma lista de reprodução, o cliente novamente estima sua capacidade de rede e faz a requisição para a próxima lista de reprodução.

2.2. Conjunto de dados

Os dados analisados neste trabalho foram retirados de registros de acesso (*logs*) gerados pelos servidores de mídia ao vivo durante a transmissão dos 64 jogos da copa do mundo de futebol da FIFA, realizado entre junho e julho de 2014.

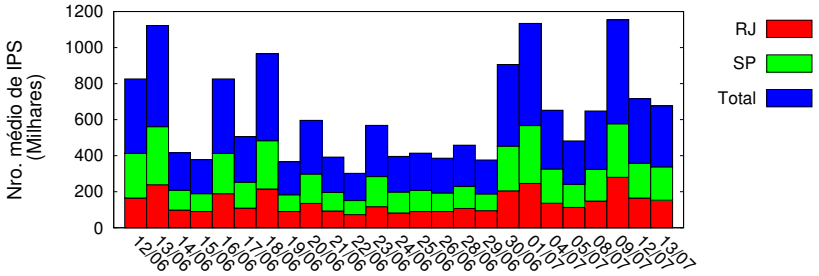


Figura 3. Média de público Internet para cada dia de transmissão da copa do mundo de futebol FIFA 2014.

Todos os jogos do evento foram transmitidos por redes de televisão, por rádio e pela Internet. Foram coletados registros de acesso da transmissão pela Internet em todos os dias em que ocorreram jogos. O conjunto de dados estudado está sumarizado na Figura 3. Em geral, os servidores atenderam entre 300 mil e 1,1 milhões de IPs únicos durante a transmissão de cada jogo, com picos de até 320 mil endereços IP únicos simultâneos. Observamos, para um único jogo, um volume total de tráfego de rede de até 300 TB e picos de até 60 GB/s.

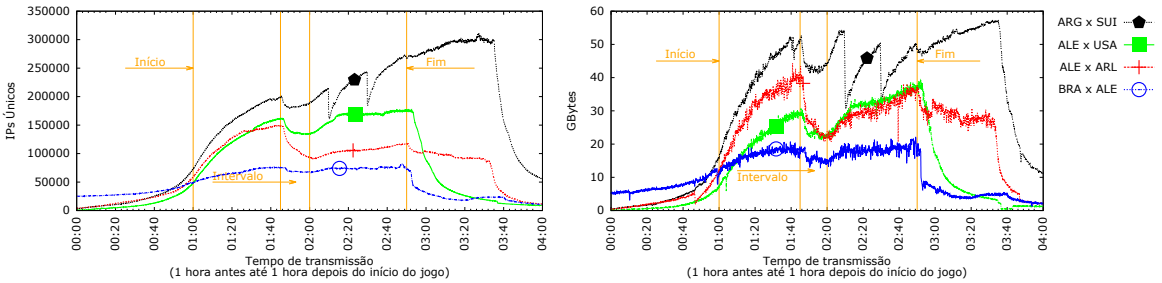
Como esperado, a carga observada durante a Copa do Mundo é maior que a carga observada durante um evento anterior: a Copa das Confederações 2013 [de Almeida Junior et al. 2015a]. O aumento de quase 4 vezes na quantidade de clientes e no volume de tráfego representou um desafio para o provedor de conteúdo naquele que foi possivelmente o evento de maior audiência no Brasil transmitido ao vivo pela Internet.

Os registros de acesso coletados dos servidores de mídia do provedor de conteúdo não incluem identificadores de sessão de usuários. Em outras palavras, apenas o endereço IP originando cada requisição é registrado. Consideramos que cada endereço IP observado corresponde a um cliente. Esta abordagem pode subestimar a quantidade de clientes em redes que utilizam NAT. Não é possível distinguir quais são as requisições de conteúdo esportivo ao vivo das requisições de mídia usuais, no entanto nossa análise não sofre impactos significativos pois elas possuem uma quantidade de tráfego desprezível.

2.3. Características gerais da carga

Diferente da copa das confederações em 2013, as configurações gerais do ambiente de transmissão de mídia ao vivo não sofreram alterações consideráveis durante os 64 jogos avaliados [de Almeida Junior et al. 2015b, de Almeida Junior et al. 2015a]. Por exemplo, durante a copa das confederações observamos alterações consideráveis das opções de qualidade de mídia sendo transmitida. Mais ainda, nossos estudos, apresentados na próxima seção, mostram que as caracterizações não revelam diferenças consideráveis no comportamento dos clientes durante uma transmissão ao vivo na copa do mundo. Por esses motivos, e por restrições de espaço, neste trabalho, focamos nossa análise em quatro jogos de interesse especial na Copa do Mundo de 2014:

- *Argentina x Suíça (Terça-feira)*: o jogo com maior carga no provedor de conteúdo;
- *Alemanha x EUA (Quinta-feira)*: o jogo mais transmitido na Internet;
- *Alemanha x Argélia (Segunda-feira)*: um jogo com alta carga e com prorrogação;
- *Alemanha x Brasil (Terça-feira)*: o fatídico jogo mais "*twittado*."



(a) Número de Usuários (b) Banda de Rede

Figura 4. Carga de trabalho a respeito dos 4 jogos de interesse.

Dentre esses jogos, apenas a partida entre Alemanha e EUA ocorre simultaneamente a outra (Portugal e Gana). A Figura 4 retrata a carga de trabalho imposta pelos quatro jogos aos servidores do provedor de conteúdo. Observe que o número de endereços IP distintos (clientes) é relativamente baixo no evento de pré-jogo, programação relativa ao jogo que se inicia cerca de uma hora antes das partidas. Entretanto, ainda no final do pré-jogo (há poucos minutos do início da partida), o número de clientes que acessa o sistema cresce de forma rápida. Geralmente, o comportamento padrão é (i) crescimento do número de clientes até minutos após o início da partida; (ii) um período razoavelmente estável durante o 1º e o 2º tempo da partida; e (iii) um intervalo caracterizado por uma queda e posterior crescimento do número de clientes. Este comportamento é estendido em jogos com prorrogação, que atraem clientes por um período de tempo mais longo. A única partida das quatro analisadas que diverge do comportamento padrão acontece entre Argentina x Suiça. Acreditamos que isso tenha ocorrido por ajuste no servidor por parte do provedor de conteúdo ao perceber o crescimento considerável no número de requisições. Nesse período há um crescimento de 100% nas requisições com erro 404.

Observamos também que o dia da semana de realização do jogo é um fator de grande impacto na atração de usuários. Esse impacto fica mais claro ainda na Figura 5, onde apresentamos a taxa de novas sessões de clientes por segundo, avaliada no período do pré-jogo e do 1º tempo da partida. Em geral, partidas aos fins de semana e feriados nacionais impõem carga menor aos servidores, possivelmente porque usuários assistem às partidas em televisores convencionais. A curva relativa à distribuição acumulada de probabilidades de um jogo realizado durante um final de semana é claramente deslocada para a esquerda. Enquanto a mediana de partidas durante a semana se encontra próximo de 90 sessões novas por segundo, a mediana da taxa de chegada de sessões de partidas durante o final de semana se encontra próximo de 40 por segundo.

Durante os jogos da copa do mundo, encontramos três níveis principais de taxa de transmissão (qualidade) de mídia pelos servidores. De acordo com a Figura 6, que apresenta a distribuição acumulada da taxa de transmissão de mídia em cada servidor separada por ISP[2], aproximadamente 40% dos usuários assistem com uma taxa de transmissão entre 150 e 400 KBps (qualidade baixa), aproximadamente 50% assistem com uma taxa de transmissão de 500 KBps (qualidade média), e aproximadamente 10% assistem com taxas de transmissão superiores a 800 KBps (qualidade alta). Usuários da *Brasil-Telecom* são

[2]Detalhes sobre localização de IPs em [de Almeida Junior et al. 2015a, de Almeida Junior et al. 2015b]

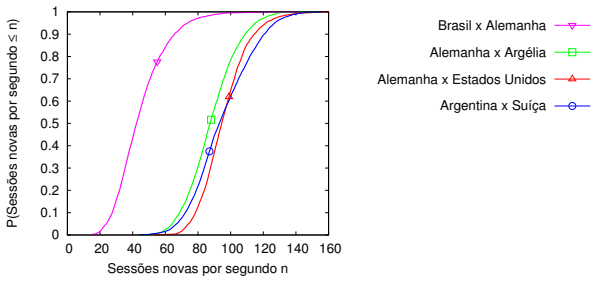


Figura 5. Taxa de criação de novas sessões em jogos representativos.

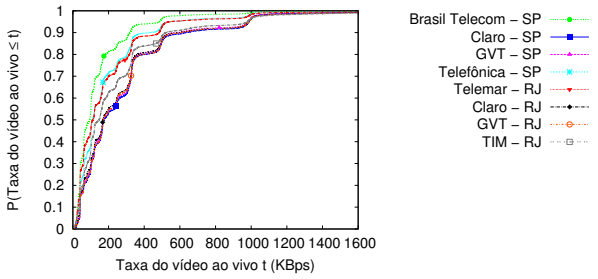


Figura 6. Taxa de transmissão da mídia ao vivo em jogos representativos.

os que assistem à mídia com a menor qualidade, enquanto usuários da Tim-RJ assistem à mídia com maior qualidade.

Finalmente, como descrevemos na seção 2.1, usuários periodicamente requisitam uma lista de reprodução. Diferente da transmissão da Copa das Confederaçõeso [de Almeida Junior et al. 2015a], as configurações para transmissão da Copa do Mundo fizeram com que os tempos entre as requisições de listas de reprodução fossem reduzidos. Verificamos que o intervalo entre as requisições de arquivos de lista de reprodução, como mostrado na Figura 7, foi bem menor do que o observado previamente, não passando de 8 segundos em 99% dos casos (o anterior chegava a 30 segundos).

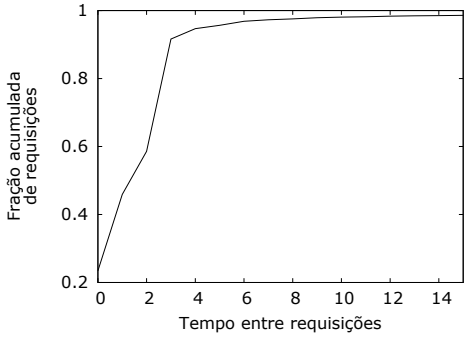


Figura 7. Distribuição do tempo entre requisições de listas de reprodução.

3. Comportamento dos clientes

Nesta seção apresentamos um modelo do comportamento dos clientes durante transmissão de vídeo ao vivo na Internet. Para cada componente do modelo, identificamos

a distribuição estatística que melhor se ajusta aos dados, dentre distribuições estatísticas amplamente usadas na literatura: normal, log-normal, exponencial, Gamma, Logística, Beta, Uniforme, Weibull e Pareto para variáveis contínuas; Poisson, Binomial, Binomial Negativa, Geométrica e Hipergeométrica para variáveis discretas. Para cada componente do modelo, os parâmetros da distribuição que mais se aproxima dos dados são determinados usando o método de estimativa por máxima verossimilhança. Após definição dos parâmetros de cada componente modelo, a distribuição com menor distância de Kolmogorov-Smirnov (distribuições contínuas) ou menor erro quadrático mínimo (LSE) (distribuições discretas) em relação aos dados é escolhida. Esta escolha também é validada com uma avaliação visual do ajuste das curvas.

3.1. Modelo do comportamento do cliente

Cada cliente estabelece uma *sessão* com os servidores de vídeo para receber e reproduzir a mídia, como mostrado na Figura 8. Durante uma sessão, o cliente requisita periodicamente a um dos servidores do provedor de conteúdo uma lista de reprodução (*pl*) indexando segmentos de mídia na codificação correspondente à qualidade que as condições de rede do cliente permitem exibir. O período de requisições de listas de reproduções e o tamanho dos segmentos de mídia são dependentes da configuração dos servidores e independem do comportamento do cliente.

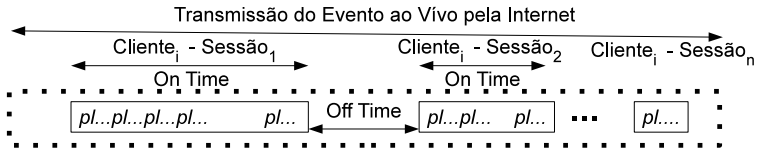


Figura 8. Comportamento de um cliente durante uma transmissão.

Cada cliente é livre para começar e interromper a visualização da mídia ao vivo a qualquer momento. A figura 8 ilustra que um cliente pode apresentar múltiplas sessões durante a transmissão de um jogo. Cada sessão tem uma duração bem definida, delimitada pelo momento em que ocorre a primeira e a última requisição a segmentos de mídia recebidas pelo provedor de conteúdo. Sessões consecutivas são espaçadas por um período compreendido entre a última requisição da sessão anterior e a primeira requisição da próxima. Por simplificação, denominamos essa duração como ON e, de forma análoga, períodos entre sessões consecutivas, denominamos por OFF.

Nós capturamos o comportamento de clientes usando um modelo "ON/OFF", como mostrado na figura 9. Neste modelo, cada cliente alterna entre o estado ativo (ON) e inativo (OFF). Durante o estado ON, o cliente está assistindo uma sessão de vídeo ao vivo. Durante o período que permanece nesse estado, cada cliente troca informações de controle com o servidor de mídia e recebe segmentos de mídia para reprodução. Após o fim de uma sessão, usuários podem voltar ao sistema de transmissão de mídia ao vivo. Nesse sentido, duas sessões ON são obrigatoriamente intercaladas por um estado de OFF. No modelo proposto, cada cliente cria uma próxima sessão, intercalada por um período OFF, com probabilidade P_{off} e pode abandonar o sistema com probabilidade $1 - P_{\text{off}}$.

O conjunto de dados que utilizamos neste trabalho não possui marcadores específicos de início e fim de sessão. Assim, nós assumimos que um usuário alterna entre um estado de atividade ON para um estado de inatividade OFF caso deixe de fazer

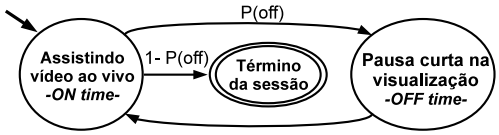


Figura 9. Modelo "ON/OFF" do comportamento de um cliente.

requisições por um período superior a um *limite de inatividade* pré-estabelecido. A escolha de um limite de inatividade apropriado é importante uma vez que valores muito pequenos podem superestimar o número de sessões, enquanto valores muito grandes podem não detectar duas sessões consecutivas que realmente existiram. Conforme apresentamos na seção 2.3, os tempos entre requisições de lista de reprodução pode ser um bom indicador desse valor limite. Nós avaliamos limites de inatividade variando entre 10 e 120 segundos. Neste intervalo, os número de sessões contabilizadas em uma partida não apresentou diferenças significativas, com diferenças médias abaixo de 10%. Dessa forma, mantivemos a metodologia de nossos trabalhos anteriores [de Almeida Junior et al. 2015a] e utilizamos um limite de inatividade de 30 segundos.

3.2. Resultados da caracterização

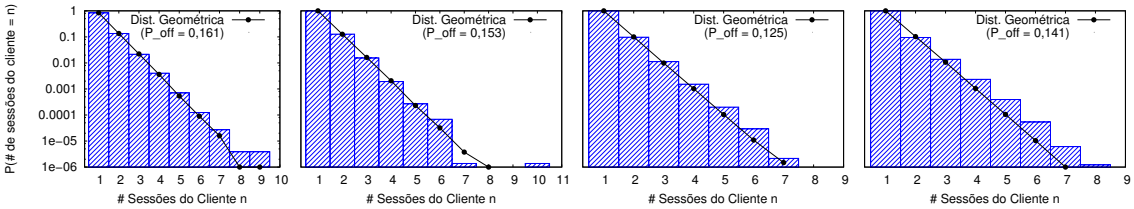
As partidas da Copa do Mundo de futebol apresentam várias nuances ao longo de seu acontecimento. Por exemplo, há um crescente número de usuários no período pré-jogo; uma saída de usuários no intervalo entre tempos e, uma saída em massa ao fim do jogo. A avaliação indiscriminada desses períodos pode adicionar ruído aos resultados. Pelo motivo exposto, nós avaliamos as características dos clientes durante um período relativamente estável. Mais precisamente, consideramos que tanto o primeiro quanto o segundo tempo da partida são relativamente estáveis, especialmente se comparados ao pré-jogo. Entretanto, escolhemos avaliar todas as amostras de sessões que se iniciaram durante o 1o. tempo da partida. Esperamos que, com essa escolha, não enviesemos os dados. Por exemplo, sessões de usuário que iniciaram próximos ao fim da partida podem gerar um número maior de sessões curtas, comparado a realidade do jogo como um todo.

A Figura 10 apresenta a distribuição do número de sessões que clientes têm ao longo de uma partida. Conforme verificamos na Figura, o comportamento entre as partidas avaliadas é muito próximo. Mais ainda, de acordo com nossas avaliações, os clientes apresentam um número pequeno de sessões. De fato, mais de 85% dos clientes tem apenas uma sessão. Apesar de existirem clientes com um número alto de sessões (e.g., 10 sessões), menos de 5% dos clientes possuem mais de duas sessões durante uma partida.

Lembre-se que, considerando o modelo ON/OFF da Figura 9, de onde nosso modelo é construído, o número de sessões por cliente pode ser computado baseando na probabilidade de existir uma transição entre o estado de atividade ON e inatividade OFF. Em outras palavras, uma distribuição Geométrica pode ser utilizada para descrever esse comportamento[3] com parâmetro $1 - P_{\text{off}}$. Na Figura 10 nós também apresentamos a distribuição geométrica de melhor ajuste aos dados apresentados. Note que as curvas e os dados medidos estão em concordância, especialmente no início da curva.

Em transmissões de vídeo ao vivo, usuários tendem a ter sessões grandes. De

[3]A função de massa de uma distribuição Geométrica com parâmetro p é $Prob(x = k) = (1 - p)^{k-1}p$.



(a) Alemanha x EUA (b) Alemanha x Argélia (c) Argentina x Suíça (d) Alemanha x Brasil

Figura 10. Número de sessões por cliente durante uma transmissão ao vivo. Uma distribuição Geométrica ajusta o nro. de sessões de um usuário. Os parâmetros dessa distribuição estão ilustrados em cada uma das figuras.

acordo com a Figura 11, praticamente 15% dos usuários tiveram sessões que cobriam toda a partida. Mais detalhadamente, as Figuras 11-a e 11-b apresentam a distribuição de probabilidade acumulada das distribuições que melhor se ajustam aos tempos ON dos 4 jogos em questão. A região em cinza, apresenta o erro padrão da média (das amostras)[4] em cada percentil da distribuição. Essa região nos dá a noção de quão distante essas distribuições estão dos dados medidos. Note que, tanto a distribuição Weibull[5], quando Gamma[6] apresentam um ajuste adequado aos dados medidos. Em ambos os casos, a diferença entre o que foi medido e o ajuste é sempre inferior a 5%.

Note também que há um número expressivo de clientes que permanecem ativos por períodos superiores ao total da partida. De fato, cerca de 10% dos tempos de ON são superiores a 100 minutos. Isso pode acontecer, principalmente, por três motivos: primeiro, os mesmos provedores responsáveis pela transmissão do conteúdo ao vivo, servem o conteúdo tradicional do provedor de serviços. Não há distinção clara para nós, nos dados coletados, a qual conteúdo cada requisição feita por clientes pertence. Assim, provavelmente há usuários assistindo a filmes ou mesmo robôs de coleta de dados em execução durante uma transmissão ao vivo da Copa do Mundo. Segundo, há transmissões relacionadas à partida mesmo após seu encerramento. Por último, entre as partidas avaliadas, há duas com prorrogação. Em especial, a partida Argentina x Suíça atraiu a atenção dos clientes de forma diferenciada.

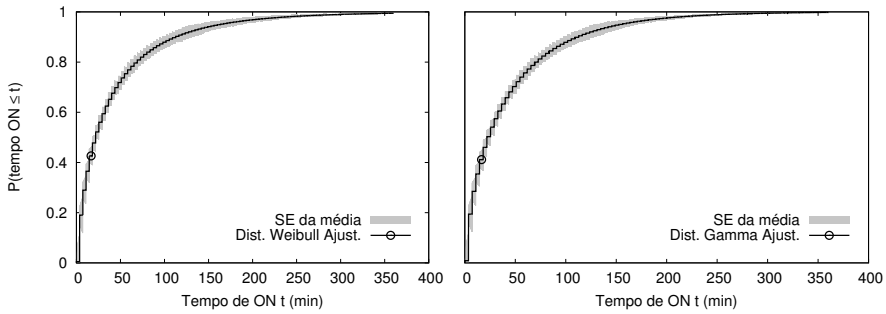
A tabela 1 detalha os parâmetros das distribuições Weibull e Gamma para cada um dos jogos avaliados, assim como o comportamento geral. Note que, a partida Argentina x Suíça destoa das demais partida. Nesse caso, enquanto a média do tempo de ON para as demais partidas ficam por volta de 34 a 38 minutos, a partida supracitada tem média de tempo de ON de 54 minutos. Além das prorrogações dessa partida, houve um crescente interesse nela, em especial, após o término do 2º tempo (Figura 4).

A Figura 12 apresenta os tempos de OFF, quando ocorrem períodos de inatividade

[4]O erro padrão da média (Standard Error of the (sample) Mean) – SEM) de um conjunto de n observações é computado como a razão entre o desvio padrão da amostra s e a raiz quadrada do número de amostras n, i.e., $s/\sqrt{n}$. Note que o $(1-\alpha)\%$ intervalo de confiança para as amostras pode ser determinado pela multiplicação do valor correspondente de SEM pelo valor do $(1-\alpha/2)$-percentil das variáveis z e t, dependendo do número de observações n [Jain 1991].

[5]Função de densidade de probabilidade (PDF) da distribuição Weibull: $p_X(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}e^{-(x/\beta)^{\alpha}}$.

[6]Gamma: $\Gamma(a) = \int_0^{\infty} s^{a-1}e^{-s}ds$.



(a) Weibull (b) Gamma

Figura 11. Tempo de duração das sessões (ON) nos quatro jogos de interesse.

Tabela 1. Distribuição do tempo de duração (ON) das sessões.

Jogo	Melhor Fitting	Média (min)	Desv. Padrão	Primeiro Parâmetro	Segundo Parâmetro
Comportamento geral	Weibull	41,93	51,94	0,700	33,708
	Gamma	41,93	51,94	0,581	0,013
Alemanha x Estados Unidos	Weibull	37,58	43,80	0,725	31,230
	Gamma	37,58	43,80	0,610	0,016
Alemanha x Argélia	Weibull	34,59	43,63	0,728	28,607
	Gamma	34,59	43,63	0,615	0,017
Argentina x Suíça	Weibull	54,33	62,66	0,743	45,987
	Gamma	54,33	62,66	0,629	0,011
Brasil x Alemanha	Weibull	38,43	51,65	0,598	26,416
	Gamma	38,43	51,65	0,472	0,012

por parte de um cliente. Assim como na Figura 11, nós apresentamos a distribuição com melhor ajuste para os dados medidos e o erro envolvido. Pela definição de atividade, que utilizamos nesse trabalho, não há tempos de OFF menores que 3 minutos. Uma distribuição LogNormal[7] é capaz de ajustar os dados medidos, com baixo erro. Note que, a grande maioria dos tempos de OFF são inferiores a 15 minutos. Mais precisamente, cerca de 85% dos tempos de OFF ocorrem por períodos que são inferiores ao tempo de intervalo que acontece no meio da partida. Intuitivamente, esse comportamento corrobora com o comportamento da carga nos servidores, observado nas Figuras 4-a e 4-b. De acordo com essas figuras, no início do intervalo nós observamos uma saída de clientes do sistema e, antes do início do 2º tempo da partida, há novamente um crescimento do número de clientes no sistema.

Finalmente, a tabela 2 resume as distribuições de cada um dos jogos, assim como a distribuição do comportamento geral. Em todas as partidas, a média dos tempos de OFF são similares. Há uma leve diferença na partida entre Alemanha x Brasil. Nesse caso, lembramos que essa partida ocorreu em um feriado nacional.

3.3. Gerador de carga sintética

Nós desenvolvemos um gerador de cargas sintéticas baseado no modelo proposto[8]. Este gerador tem por objetivo produzir um registro de acessos de usuários a um sistema de transmissão de mídia ao vivo, imitando o comportamento dos clientes reais durante um

[7]PDF da distribuição Log-Normal: $p_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(ln(x)-\mu)^2}{2\sigma^2}}$.

[8]Disponível em http://netlab.ice.ufjf.br.

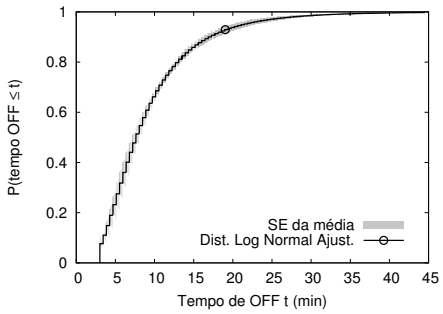


Figura 12. Tempo entre sessões (OFF) nos quatro jogos de interesse.

Tabela 2. Distribuição do tempo entre sessões (OFF).

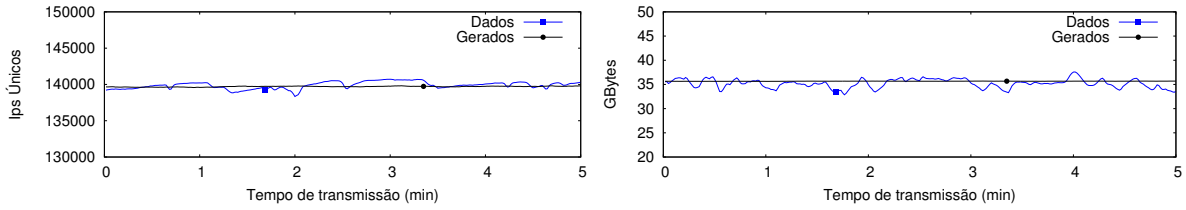
Jogo	Melhor Fitting	Média (min)	Desv. Padrão	Primeiro Parâmetro	Segundo Parâmetro
Comportamento geral	LogNormal	9,28	6,86	2,010	0,634
Alemanha x Estados Unidos	LogNormal	9,21	6,94	1,996	0,643
Alemanha x Argélia	LogNormal	9,63	6,92	2,053	0,632
Argentina x Suíça	LogNormal	9,33	6,90	2,014	0,637
Brasil x Alemanha	LogNormal	8,69	6,47	1,956	0,608

grande evento. Note que, não pretendemos nesse momento, gerar comportamentos que são dependentes do servidor. Assim, a grande variabilidade de carga notada durante uma partida não é objetivo desse simulador. Em particular, a partir de um desejado número de clientes, o gerador cria o comportamento dinâmico de cada usuário do sistema a partir das distribuições do número de sessões por partida, duração de sessão (tempo ON) e tempo entre sessões (tempo OFF). Como trabalho futuro, consideramos estender o gerador para capturar o tráfego sob a perspectiva de cada usuário. Neste caso, a escolha da qualidade da mídia e outros parâmetros dependentes do provedor de conteúdo deve ser modelada de forma mais detalhada.

Nós validamos o gerador de forma direta e indireta. Por restrições de espaço, exibiremos apenas a segunda. Ela é mais interessante pois foca em métricas que não são explicitamente modeladas pelo gerador. Para tal validação, consideramos o número de usuários de um sistema e a taxa da mídia sendo transmitida.

O gerador produziu um registro sintético com aproximadamente n usuários conectados ao sistema. Para cada um, o gerador criou o número de sessões correspondente, seus tempos de ON e de OFF. Para cada instante de tempo, nós contabilizamos o número de clientes ativos. Nós associamos uma taxa média de mídia para cada um dos n clientes ativos e contabilizamos a banda de rede para aquele momento de tempo. Nós escolhemos um período estável de um jogo real para comparar com o resultado do simulador. Escolhemos o período do minuto 25-30, do jogo Argentina x Argélia, com cerca de 140 mil clientes (figura 4). Note que, o simulador não é limitado nem direcionado a esse período.

A figura 13 apresenta a população de usuários criadas pelo gerador e a taxa de transmissão que essa população consome do provedor de conteúdo. Idealmente, o gráfico deveria apresentar os pontos próximos ao dados medidos, incluindo sua alta variabilidade. Entretanto, isto é muito difícil de se alcançar, principalmente quando o processo sendo modelado é complexo. Em particular, nosso gerador não considera variação do número



(a) Ips únicos (b) Volume (GB)

Figura 13. Validação do Gerador de Cargas Sintéticas.

de clientes nem da taxa de transmissão da mídia (figura 6), apenas usamos a média. Em suma, apesar das imprecisões, consideramos que o gerador captura o nível de carga e padrões de acesso a servidores do provedor de conteúdo de mídia ao vivo, incluindo seus tempos de atividade e inatividade.

4. Discussões e Implicações

Claramente, provedores atuais ainda não estão satisfatoriamente preparados para uma nova era onde eventos de larga escala serão regra e não mais exceção. Por exemplo, a Copa do Mundo de futebol em 2014, notoriamente um evento de grande escala, movimentou todo o Brasil. Em especial, empresas de TI se prepararam para atender a demanda por comunicação na Internet. Mesmo com todo o preparo, grandes empresas de conteúdo foram surpreendidas por demandas muito elevadas. Por exemplo, a ESPN teve picos de 1,7 milhões clientes simultâneos durante o jogo mais transmitido pela Internet até a atualidade. Toda sua infraestrutura de redes e de comunicação sofreu com essa demanda e a companhia ficou sem conexão durante grande parte da partida.[9]

Grandes eventos na Internet costumam ser bem planejados, com meses de preparação antecedente. Parte da estrutura é testada por cargas sintéticas, que tentam imitar o comportamento esperado de um cliente para aquele serviço. Em especial, para transmissão de vídeo na Internet e conteúdo web, há uma série de trabalhos que se dedicam a geração de carga sintética [Jin and Bestavros 2001, Busari and Williamson 2002, Tang et al. 2003, Krishnamurthy et al. 2006]. Apesar da importância histórica desses trabalhos, eles foram realizados em um período ainda de surgimento de aplicações com alta demanda de carga na Internet. Naquela época ainda havia poucas caracterizações acerca do comportamento de usuários e, diversas características comuns nos dias de hoje sequer existiam.

As principais caracterizações sobre o comportamento de usuários em transmissões ao vivo na Internet datam de cerca de uma década atrás. Por exemplo, [Veloso et al. 2006] avaliaram o comportamento de clientes em uma transmissão contínua de um dos primeiros *reality show* transmitido no Brasil. Nesse trabalho os autores propõem um modelo para o comportamento dos clientes e caracterizam seus componentes. Transmissões ao vivo em arquiteturas P2P se tornaram populares a partir da segunda metade da década de 2000. Nessa mesma época, surgiram caracterizações de eventos transmitidos nesse tipo de aplicação. Por exemplo, [Hei et al. 2007] estudam um sistema de transmissão ao vivo em P2P e avaliam características de seus clientes durante uma transmissão de um grande evento na China. Mais recentemente, [Borges et al. 2012] propõem um modelo

[9]http://www.businessinsider.com/espn-down-during-world-cup-2014-6

hierárquico para descrever o comportamento dos clientes no SopCast, uma aplicação popular de transmissão ao vivo em P2P. Os autores diferem o comportamento de clientes de transmissões de eventos recorrentes (e.g., novelas, jornais) e eventos únicos (e.g., finais de campeonato de jogos de futebol).

Nosso trabalho, diferente dos existentes na literatura, é focado em uma era onde transmissão ao vivo está consolidada. Nós focamos em eventos de larga escala, uma vez que tais eventos impactam os provedores de conteúdo de forma marcante. Além disso, nosso trabalho evidencia as mudanças observadas no comportamento dos clientes ao longo do tempo, deixando claro que modelos e caracterizações existentes podem estar obsoletas e, por consequência, não são apropriadas para serem utilizadas atualmente. Listamos as seguintes diferenças marcantes no comportamento dos clientes, confrontados com os principais trabalhos existentes:

- Número de sessões: o número médio de sessões de um determinado usuário está ainda menor. De fato, em transmissões de eventos em rede P2P, 40% dos clientes tinham uma única sessão [Borges et al. 2012]. Agora, cerca de 80% dos clientes têm apenas uma única sessão.
- Tempo de ON: o tempo de ON está mais longo, com médias de aproximadamente 40 min. De fato, observamos na prática a substituição de sistemas de TV tradicional por aplicativos de Internet para visualização de programação ao vivo de emissoras. Isso tem aumentado a duração de tempo de sessões dos usuários de vídeo ao vivo na Internet. Comparado a trabalhos anteriores, esse aumento é significativo. Por exemplo, durante uma partida de final de campeonato de futebol, cerca de 60% dos clientes tiveram um tempo muito curto de ON (menos de 3 min) e apenas 5% assistiram quase toda a partida [Borges et al. 2012]. Nessa mesma linha, [Hei et al. 2007] apontam que 90% das sessões duram menos que 10 min.
- Tempo de OFF: os tempos de inatividade também estão alterados atualmente. Enquanto em [Borges et al. 2012], 35% dos clientes apresentavam tempos de OFF superiores a 20 minutos, no trabalho atual, apontamos que apenas cerca de 6% dos usuários ficam inativos pelo mesmo período.

5. Conclusões e Trabalhos Futuros

Neste artigo, apresentamos uma caracterização e modelagem dos clientes de um sistema de vídeo ao vivo durante um evento de larga escala na Internet. O estudo foi feito a partir de cargas de trabalho da transmissão das 64 partidas de futebol da Copa do Mundo de 2014, fornecidas por um dos principais provedores de conteúdo do Brasil. Para analisar o comportamento dos clientes foi proposto um modelo simples que captura as principais características de um cliente, como sua forma de interação com os provedores de conteúdo, seu tempo de sessão e seus períodos de inatividade.

Quanto ao comportamento dos clientes, observamos que há pouca variação entre os jogos avaliados, o que sugere que as distribuições encontradas descrevem bem seus comportamentos. A maioria das sessões são longas, e a grande maioria dos clientes apresenta uma única sessão. Além da diferença das distribuições que descrevem alguns comportamentos dos clientes, encontramos também diferenças qualitativas significativas marcantes. Por exemplo, o tempo de sessão é notoriamente maior durante os eventos atuais, assim como o tempo de inatividade, quando existente, se reduziu. Isso indica que os

padrões de carga de eventos são significativamente distintos dos existentes na literatura.

Trabalhos futuros incluem extensões para novos conjuntos de dados e a extensão do gerador de cargas sintéticas para outros aspectos da carga de trabalho, por exemplo, taxa de chegada de clientes e adaptação da taxa de transmissão da mídia. Mais ainda, dada a natureza dos eventos, com grandes volumes de dados, pretendemos criar mecanismos para descrever o comportamento da carga de trabalho a partir de um número pequeno de variáveis.

Referências

Barroso, L. A., Clidaras, J., and Hölzle, U. (2013). The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis lectures on computer architecture*, 8(3):1–154.

Borges, A., Gomes, P., Nacif, J., Mantini, R., de Almeida, J. M., and Campos, S. (2012). Characterizing SopCast Client Behavior. *Comput. Commun.*, 35(8):1004–1016.

Busari, M. and Williamson, C. (2002). ProWGen: A Synthetic Workload Generation Tool for Simulation Evaluation of Web Proxy Caches. *Comput. Netw.*, 38(6):779–794.

Cesario, M. V. (2012). Uso de anycast para balanceamento de carga na globo.com. *Talks and Tutorial, SBRC 2012*. Disponível em: `http://pt.slideshare.net/marcuscesario/apresentacao-anycast-sbrc201205`.

de Almeida Junior, W., Almeida, B., Ítalo Cunha, Almeida, J. M., and Vieira, A. B. (2015a). Caracterização do Tráfego e Impacto de Rede da Transmissão de um Grande Evento Esportivo. In *33o. Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos - SBRC*.

de Almeida Junior, W., Borges, A. V., and Ítalo Cunha (2015b). Avaliação de transmissão ao vivo de grandes eventos pela internet. *PPGCC-UFJF - Dissertação de mestrado*. Disponível em: `http://www.ufjf.br/pgcc/files/2014/06/WagnerAlmeida.pdf`.

Gebert, S., Pries, R., Schlosser, D., and Heck, K. (2012). *Internet access traffic measurement and analysis*. Springer.

Hei, X., Liang, C., Liang, J., Liu, Y., and Ross, K. W. (2007). A measurement study of a large-scale p2p iptv system. *Multimedia, IEEE Transactions on*, 9(8):1672–1687.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley & Sons.

Jin, S. and Bestavros, A. (2001). GISMO: A Generator of Internet Streaming Media Objects and Workloads. *SIGMETRICS Perform. Eval. Rev.*, 29(3):2–10.

Katabi, D. and Wroclawski, J. (2000). A framework for scalable global ip-anycast (gia). *ACM SIGCOMM Computer Communication Review*, 30(4):3–15.

Krishnamurthy, D., Rolia, J. A., and Majumdar, S. (2006). A Synthetic Workload Generation Technique for Stress Testing Session-Based Systems. *IEEE Trans. Softw. Eng.*, 32(11):868–882.

Tang, W., Fu, Y., Cherkasova, L., and Vahdat, A. (2003). MediSyn: A Synthetic Streaming Media Service Workload Generator. In *Proc. of the NOSSDAV*, pages 12–21.

Veloso, E., Almeida, V., Jr., W. M., Bestavros, A., and Jin, S. (2006). A Hierarchical Characterization of a Live Streaming Media Workload. *IEEE/ACM Trans. Netw.*, 14(1):133–146.